North-Holland

Resistant lower rank approximation of matrices by iterative majorization [1]

Peter Verboon and Willem J. Heiser
University of Leiden, Leiden, The Netherlands

Received September 1992
Revised May 1993

Abstract: It is commonly known that many techniques for data analysis based on the least squares criterion are very sensitive to outliers in the data. Gabriel and Odoroff (1984) suggested a resistant approach for lower rank approximation of matrices. In this approach, weights are used to diminish the influence of outliers on the low-dimensional representation. The present paper uses iterative majorization to provide for a general algorithm for such resistant lower rank approximations which guarantees convergence. It is shown that the weights can be chosen in different ways corresponding with different objective functions. Some possible extensions of the algorithm are discussed.

Keywords: Lower rank approximation; Resistance; Robustness; Huber function; Biweight function; Iteratively reweighted least squares; Majorization.

1. Introduction

Finding a low-dimensional representation of a high-dimensional data matrix is a well-known method in data analysis. For instance, the biplot (Gabriel, 1971) and principal component analysis (PCA) are among the most important multivariate techniques in data exploration. The criterion used to examine how well the data are represented, is usually defined in terms of the squared residuals. However, when the least squares criterion is used and there are outliers in the data, the low-dimensional representation may not be the most interesting one, and will tend to be unstable. This problem has been investigated by Hawkins and Fatti (1984) for outliers that deflate the correlations and dominate the last few principal components. A general discussion of the influence of outliers in PCA is given in Joliffe (1986, ch. 9).

Correspondence to: Peter Verboon, Department of Data Theory, Faculty of Social Sciences, University of Leiden, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

[1] This research was supported by a PSYCHON grant (560-267-029) of the Netherlands organization for scientific research (NWO) for the first author.

SSDI 0167-9473(93)E0027-2

Gabriel and Zamir (1979) show how a weighted least squares algorithm can be used to find a low-dimensional representation of both column and row points of a data matrix. In this approach each entry in the data matrix is separately weighted with some prechosen nonnegative quantity. An extension of this basic idea is given by Gabriel and Odoroff (1984) who suggest to use these weights to decrease the influence of outliers on the representation. In this extension, the weights are related to the residuals from the low-dimensional representation, which yields an iterative scheme, known as iteratively reweighed least squares (IRLS), in which the representation and the weights are alternatingly updated.

In the present paper we will use a majorization argument to prove convergence of IRLS algorithms that obtain a low-dimensional representation which is not influenced by outliers. It will be shown that iterative majorization can be used for a variety of resistant loss functions, by merely choosing the weights differently. This makes the iteratively reweighted Gabriel-Zamir algorithm widely applicable.

2. The method of successive dyadic fitting

Let $\mathbf{Z} = \{z_{ij}\}$ be the observed data matrix of order $n \times m$. A p-dimensional ($p \leq m$) representation of $\mathbf{Z}$ is given by

$$\mathbf{Z} \cong \mathbf{X}\mathbf{A}', \tag{1}$$

where $\cong$ represents the least squares approximation. The row markers or component (object) scores are in the matrix $\mathbf{X}$ of order $n \times p$ ($p \leq m$), and the matrix $\mathbf{A}$ ($m \times p$) contains the column markers or component loadings.

Finding $\mathbf{X}$ and $\mathbf{A}$ in (1) implies that we must minimize the following loss function:

$$\sigma(\mathbf{X}, \mathbf{A}) = \mathrm{tr}(\mathbf{Z} - \mathbf{X}\mathbf{A}')'(\mathbf{Z} - \mathbf{X}\mathbf{A}'), \tag{2}$$

with the normalization constraint $\mathbf{X}'\mathbf{X} = n\mathbf{I}_p$. The normalization constraint is necessary for identification, since the product $\mathbf{X}\mathbf{A}'$ is unique up to linear transformations of $\mathbf{X}$ and $\mathbf{A}$. First consider a $p = 1$ approximation of $\mathbf{Z}$. In this case $\mathbf{X}$ reduces to the column vector $\mathbf{x}_1$ and $\mathbf{A}$ to the column vector $\mathbf{a}_1$. The loss function for the first principal component then becomes

$$\sigma(\mathbf{x}_1, \mathbf{a}_1) = \mathrm{tr}(\mathbf{Z} - \mathbf{x}_1\mathbf{a}_1')'(\mathbf{Z} - \mathbf{x}_1\mathbf{a}_1'). \tag{3}$$

Both unknown vectors $\mathbf{x}_1$ and $\mathbf{a}_1$ are easily computed via simple regression equations (Good, 1969), which is called dyadic fitting (Gabriel & Zamir, 1979). For the first component we may iterate between $\mathbf{x}_1$ and $\mathbf{a}_1$ until the solution stabilizes. After convergence the product $\mathbf{x}_1\mathbf{a}_1'$ is the best least squares rank-one approximation of $\mathbf{Z}$.

Subsequent dimensions can be found as follows. The data are replaced by the

residuals from the rank-one approximation by subtracting the approximation from the original data; thus

$$\mathbf{Z}^{(-1)} = \mathbf{Z} - \mathbf{Z}^1,$$

where $\mathbf{Z}^1$ represents the rank-one approximation. With this $\mathbf{Z}^{(-1)}$, (3) can be solved again for the second dimension, which yields new $\mathbf{x}_2$, $\mathbf{a}_2$, and $\mathbf{Z}^{(-2)}$. In this way all p dimensions (principal components) can be computed. This stepwise fitting is possible because successive columns of $\mathbf{X}$, and also successive columns of $\mathbf{A}$, are orthogonal (or, in the case of multiple singular values of $\mathbf{Z}$, they can be chosen to be orthogonal).

3. Weighted cyclic dyadic fitting

Let $\mathbf{W} = \{w_{ij}\}$ be a matrix with weights of order $n \times m$, so that each w_{ij} corresponds with an observation z_{ij} in the data. Furthermore, let $\mathbf{V}_j$ $(j = 1, \ldots, m)$ be a diagonal matrix with the elements w_{ij} $(i = 1, \ldots, n)$ for some j on the diagonal. The weighted least squares loss function is now written as

$$\sigma(\mathbf{X}, \mathbf{A}) = \sum_{j=1}^{m} (\mathbf{z}_j - \mathbf{X}\mathbf{a}_j)' \mathbf{V}_j (\mathbf{z}_j - \mathbf{X}\mathbf{a}_j), \tag{4}$$

where $\mathbf{a}_j$ and $\mathbf{z}_j$ are the jth column of $\mathbf{A}'$ and $\mathbf{Z}$, respectively. It is not hard to see that minimizing (4) over $\mathbf{X}$ and $\mathbf{A}$ can be done by alternatingly solving a weighted least squares problem. For given $\mathbf{Z}$ and $\mathbf{X}$ it follows directly from (4) that each $\mathbf{a}_j$ is found independently by projecting $\mathbf{z}_j$ on the space spanned by the columns of $\mathbf{X}$ in the metric $\mathbf{V}_j$, giving regression weights $\mathbf{a}_j$. Rewriting (4) as a summation over rows shows that for given $\mathbf{Z}$ and $\mathbf{A}$ the rows of $\mathbf{X}$ are also regression weights that are found by projecting $\mathbf{z}_i$ on the space spanned by the columns of $\mathbf{A}$ in the metric $\mathbf{V}_i$, since

$$\sigma(\mathbf{X}, \mathbf{A}) = \sum_{i=1}^{n} (\mathbf{z}_i - \mathbf{A}\mathbf{x}_i)' \mathbf{V}_i (\mathbf{z}_i - \mathbf{A}\mathbf{x}_i), \tag{5}$$

where $\mathbf{V}_i (m \times m)$ is diagonal with the elements of the ith row of $\mathbf{W}$ on the diagonal; $\mathbf{x}_i$ and $\mathbf{z}_i$ are the ith row of $\mathbf{X}$ and $\mathbf{Z}$, respectively. It is important to note that this weighted alternating least squares procedure can no longer be applied successively, but must be carried out cyclically (Gabriel & Zamir, 1979). Thus, we may start as in the unweighted case by first computing the solution for the first dimension, subtracting this solution from the data and continue with these residualized values to compute the next dimension. But we have to cycle through this process again. In general, the elements a_{jk} are updated by

$$a_{jk} = (\mathbf{x}_k' \mathbf{V}_j \mathbf{x}_k)^{-1} \mathbf{x}_k' \mathbf{V}_j \tilde{\mathbf{z}}_j, \tag{6}$$

and the elements x_{ik} by

$$x_{ik} = (\mathbf{a}_k' \mathbf{V}_i \mathbf{a}_k)^{-1} \mathbf{a}_k' \mathbf{V}_i \tilde{\mathbf{z}}_i \tag{7}$$

where $\mathbf{x}_k$ and $\mathbf{a}_k$ are the kth column of $\mathbf{X}$ and $\mathbf{A}$, respectively, and $\tilde{\mathbf{z}}_j$ and $\tilde{\mathbf{z}}_i$ are the jth column and the ith row of the residualized matrix

$$\tilde{\mathbf{Z}} = \mathbf{Z} - \sum_{l \neq k} \mathbf{x}_l \mathbf{a}_l', \tag{8}$$

in which the contribution of the dyadic fits of the other dimensions have been substracted from the data matrix. Cycling is necessary because the succesive $\mathbf{a}_k$'s (as well as the $\mathbf{x}_k$'s) are generally not orthogonal except when the weights are equal.

4. Resistant loss functions

The vulnerability of the least squares criterion in the presence of outliers is well-known. In the context of estimating a location parameter and in regression analysis alternative loss functions have been introduced, such as Huber's function (Huber, 1964; 1981) and Tukey's biweight function (Beaton & Tukey, 1974). These functions have proved to be a good alternative for least squares when there are outliers in the data. In the present paper we will show that these functions can also be applied in the situation where a low-dimensional approximation of a data matrix is required. To formulate the problem in a general way, we rewrite it as a summation over residual elements

$$\sigma(\mathbf{X}, \mathbf{A}) = \sum_{j=1}^{m} \sum_{i=1}^{n} f(r_{ij}),$$

where the residuals are defined by $r_{ij} = z_{ij} - \sum_{k=1}^{p} x_{ik} a_{jk}$. The ordinary least squares function is of course given by $f(r_{ij}) = r_{ij}^2$. The Huber function, for each separate residual element, is defined as

$$f_{\mathrm{H}}(r_{ij}) = \begin{cases} \frac{1}{2} r_{ij}^2 & \text{if} \quad |r_{ij}| < c, \\ c\,|r_{ij}| - \frac{1}{2}c^2 & \text{if} \quad |r_{ij}| \geq c, \end{cases} \tag{9}$$

where the constant c is called the tuning constant. For small residuals the ordinary least squares function is used, while for relatively large residuals the least absolute residuals criterion is inserted. This differential treatment of residuals implies that with the Huber function the influence of large deviations is reduced compared to least squares. If c is made very small, so that all residuals are larger than than c, minimizing the Huber function amounts to minimizing the L_1 norm. The biweight function is even more radical in down weighting the large residuals:

$$f_{\mathrm{B}}(r_{ij}) = \begin{cases} \frac{1}{6}c^2\left(1 - \left(1 - \left(r_{ij}/c\right)^2\right)^3\right) & \text{if} \quad |r_{ij}| \leq c, \\ \frac{1}{6}c^2 & \text{if} \quad |r_{ij}| > c. \end{cases} \tag{10}$$

Tukey's biweight is called a hard redescending function, because its first

derivative first ascends and then descends, while it becomes exactly zero for large residuals, which implies that it has a relatively high tolerance towards large deviations and that it is indifferent beyond c. It follows that outliers in the data may become associated with large residuals. They can be arbitrarily far away from the model, since their contribution to the loss is constant. Consequently they have no further influence upon the solution, which is presumably entirely based on "good" points only.

5. Minimization by iterative majorization

To minimize the Huber and biweight functions, the iterative majorization method will be used. This method was first explicitly used in this context by Heiser (1987) and recently in a slightly different context by Verboon and Heiser (1992). In the present paper it will be shown that majorization leads to Gabriel and Odoroffs (1984) iteratively reweighted least squares (IRLS) algorithm, with two main steps. In one step, a weighted least squares problem is solved for a fixed set of weights, and in the other, the weights are chosen as a monotonically decreasing function of the residuals from the previous step.

The principle of iterative majorization relies on a family of functions $\mu(\cdot)$ for which the following inequality holds

$$\mu(r_{ij}; w_{ij}) \geq f(r_{ij}). \tag{11}$$

The notation $\mu(r_{ij}; w_{ij})$ says that $\mu(r_{ij}; w_{ij})$ is a function of the residuals r_{ij} for some set of fixed weights w_{ij} based on residuals (r_{ij}^*) that have been derived in the previous step of the algorithm. The majorizing function $\mu(r_{ij}; w_{ij})$ should be chosen in such a way that this function is much easier to minimize than the loss function itself. At each step in the algorithm $\mu(r_{ij}; w_{ij})$ is adapted, using the residual of the previous step as a so-called supporting point. To identify $\mu(r_{ij}; w_{ij})$ the following equality should also hold

$$\mu(r_{ij}^*; w_{ij}) = f(r_{ij}^*). \tag{12}$$

Together with (11), equality (12) implies that at r_{ij}^* both functions have the same first derivative (if it exists). If this derivative is not zero (in which case the minimum would be attained), then we can always find new residuals (r_{ij}^+), which minimize $\mu(r_{ij}; w_{ij})$, such that

$$\mu(r_{ij}^+; w_{ij}) < \mu(r_{ij}^*; w_{ij}). \tag{13}$$

The updated residual will be used as the new supporting point, except when $\mu(r_{ij}^+; w_{ij}) = \mu(r_{ij}^*; w_{ij})$, in which case the algorithm stops. Combining (11), (12), and (13) yields $f(r_{ij}^+) \leq f(r_{ij}^*)$ with equality only at the minimum, which implies that each step decreases the value of the objective function.

For this argument to apply in the case of resistant loss functions, it must be verified that there exist majorizing functions for $f_H(r_{ij})$ and $f_B(r_{ij})$ defined in

(9) and (10). For instance, a majorizing function $\mu_B(\cdot)$ for minimizing the loss components in the biweight, can be defined as

$$\mu_B(r_{ij}; w_{ij}) = c^2/6\left(1 - 3w_{ij}\left(1 - \left(r_{ij}/c\right)^2\right) + 2w_{ij}^{3/2}\right). \tag{14}$$

From (14) it is clear that this majorizing function is a weighted quadratic function of the residuals. After dropping all irrelevant constants, this yields the same minimization problem as the one given in (4) and (5). It follows that in the IRLS algorithm we can minimize (4) and (5) for some fixed weights by applying alternating least squares, and in the other step we update the weights.

Different choices for the weights function correspond to different resistant functions. For the Huber function the weights have to be computed as

$$w_{ij} = \begin{cases} 1 & \text{if} \quad |r_{ij}^*| < c, \\ \dfrac{c}{|r_{ij}^*|} & \text{if} \quad |r_{ij}^*| \geq c, \end{cases} \tag{15}$$

The weights for the biweight function are found by

$$w_{ij} = \begin{cases} \left(1 - \left(r_{ij}^*/c\right)^2\right)^2 & \text{if} \quad |r_{ij}^*| \leq c, \\ 0 & \text{if} \quad |r_{ij}^*| > c. \end{cases} \tag{16}$$

In both cases, we obtain a set of weights ($0 \leq w_{ij} \leq 1$) that is monotonically decreasing with respect to the absolute values of the previous residuals, as can easily be verified from (15) and (16), and which can be used as diagnostics. Weights close to 1 are assigned to data that fit the model well, while badly fitting points (outliers) will have small weights. A short and informal overview of the algorithm is presented in Figure 1.

It will now be shown that $\mu_B(r_{ij}; w_{ij})$ is indeed a majorizing function for the biweight function given in (10). The two conditions (11) and (12) must hold for a proper majorizing function.

Lemma 1 For the previously found parameter matrices, yielding residuals r_{ij}^*, the value of the biweight function is equal to that of function (14), i.e. $f_B(r_{ij}^*) = \mu_B(r_{ij}^*; w_{ij})$.

Proof For notational convenience, we will set $c = 1$. Considering one loss component, substitution of (16) in (14) yields for the first part of the function:

$$\begin{aligned} \mu_B(r_{ij}^*; w_{ij}) &= 1/6\left(1 - 3\left(1 - r_{ij}^{*2}\right)^3 + 2\left(1 - r_{ij}^{*2}\right)^3\right) = 1/6\left(1 - \left(1 - r_{ij}^{*2}\right)^3\right) \\ &= f_B(r_{ij}^*). \end{aligned}$$

The second part of both functions is equal to $1/6$ because $w_{ij} = 0$. Since the equality can be proved for any component, it consequently has been proved for the summation over the components too. $\therefore$

```
START
initialization

while still improvements of overall loss found do
    compute new weights w_ij using (15) or (16)
    while still improvements of loss found for fixed weights w_ij do
        for j = 1 to m do
            if measurement level of jth variable is ordinal or nominal then
                compute unrestricted update q_j+
                constrain q_j+ to q_j ε Γ_j
                set q_j equal to standardized q_j
            end if
        end for

        while still improvements for X found do
            for k = 1 to p do
                update x_k according to (7)
            end for
        end while

        for j = 1 to m do
            update a_j according to (6)
        end for
    end while
end while

STOP
```

Fig. 1. Schematic overview of iterative majorization algorithm.

Lemma 2 The value of function (14) is never smaller than the value of the biweight function, i.e. $f_B(r_{ij}) \leq \mu_B(r_{ij}; w_{ij})$.

Proof There are two situations: (i) $|r_{ij}| > 1$ and (ii) $|r_{ij}| \leq 1$. In situation (i) we have

$$\tfrac{1}{6} \leq \tfrac{1}{6}\left(1 - 3w_{ij}\left(1 - r_{ij}^2\right) + 2w_{ij}^{3/2}\right). \tag{17}$$

This inequality is true since $w_{ij}\,(1 - r_{ij}^2) \leq 0$ and $w_{ij}^{3/2} \geq 0$; thus, the term

In situation (ii) we have

$$\tfrac{1}{6}\left(1-\left(1-r_{ij}^2\right)^3\right) \leq \tfrac{1}{6}\left(1-3w_{ij}\left(1-r_{ij}^2\right)+2w_{ij}^{3/2}\right). \tag{18}$$

We start from the general inequality $(a-b)^2 \geq 0$, which gives $a^2 \geq 2ab - b^2$. Using this inequality we may also write:

$$\left(1-r_{ij}^2\right)^2 \geq 2\left(1-r_{ij}^2\right)\left(1-r_{ij}^{*2}\right) - w_{ij}. \tag{19}$$

Next both sides are multiplied by the non-negative quantity $(1-r_{ij}^2)$, yielding:

$$\left(1-r_{ij}^2\right)^3 \geq 2\left(1-r_{ij}^2\right)^2\left(1-r_{ij}^{*2}\right) - w_{ij}\left(1-r_{ij}^2\right). \tag{20}$$

Substituting the second part of (19) for the term $(1-r_{ij}^2)^2$ does not change the inequality:

$$\left(1-r_{ij}^2\right)^3 \geq 2\left[2\left(1-r_{ij}^2\right)\left(1-r_{ij}^{*2}\right) - w_{ij}\right]\left(1-r_{ij}^{*2}\right) - w_{ij}\left(1-r_{ij}^2\right). \tag{21}$$

Working out this expression yields

$$\left(1-r_{ij}^2\right)^3 \geq 3\left(1-r_{ij}^2\right)w_{ij} - 2w_{ij}^{3/2}. \tag{22}$$

Substracting both terms from one and multiplying by $1/6$ gives

$$\tfrac{1}{6}\left(1-\left(1-r_{ij}^2\right)^3\right) \leq \tfrac{1}{6}\left(1-3\left(1-r_{ij}^2\right)\left(1-r_{ij}^{*2}\right)^2+2\left(1-r_{ij}^{*2}\right)^3\right), \tag{23}$$

which proves the inequality. Again, if this inequality is true for any element, it is also true for the summation. $\therefore$

From these two lemma's it follows that $\mu_B(r_{ij};\, w_{ij})$ can be used as a majorizing function for the biweight function. Since (10) is bounded from below, majorization theory guarantees that at least a local minimum is attained by alternating repeatedly between minimizing (14) and updating the weights through (16), in case of the biweight.

For the Huber function, the simplest majorizing function is

$$\mu_{\mathrm{H}}(r_{ij};\, w_{ij}) = \begin{cases} \tfrac{1}{2}w_{ij}r_{ij}^2 & \text{if} \quad r_{ij}^* < c \\ \tfrac{1}{2}w_{ij}r_{ij}^2 + cr_{ij}^* - c^2 & \text{if} \quad r_{ij}^* \geq c. \end{cases} \tag{24}$$

which is also quadratic in the residuals. This function can be used to majorize Huber's function, because of the following lemma's:

Lemma 3 $f_{\mathrm{H}}(r_{ij}^*) = \mu_{\mathrm{H}}(r_{ij}^*;\, w_{ij})$

Lemma 4 $f_{\mathrm{H}}(r_{ij}) \leq \mu_{\mathrm{H}}(r_{ij};\, w_{ij})$

The proofs of lemma 3 and 4 can be found in Heiser (1987).

6. Aggregating the residuals

In the previous section, different weight matrices $\mathbf{V}_j$ were considered, one for each variable. An interesting special case occurs when we assume that all $\mathbf{V}_j$ are equal to each other; thus, $\mathbf{V}_j = \mathbf{V}$ for $j = 1, \ldots, m$. This implies we have n weights, one for each object, instead of $n \times m$ weights, one for each cell in the data matrix. Thus, instead of applying the loss function to each residual element, r_{ij}, and summing to obtain the overall loss, we aggregate residuals over rows and apply the loss function to these n aggregated values, denoted as d_i $(i = 1, \ldots, n)$. Handling the residuals rowwise, we should first compute the residuals per row, d_i. The value d_i is computed as the Euclidean distance between an object in the m-dimensional space and its model values that satisfy the rank-p restrictions; thus,

$$d_i = \sqrt{\sum_{j=1}^{m} \left(z_{ij} - \sum_{k=1}^{p} x_{ik} a_{jk} \right)^2} . \tag{25}$$

Now, the Huber or biweight function can be applied to these values to obtain a loss per row; a summation of these row losses yields the total loss.

For least squares both ways of handling the residuals are equivalent, but for the Huber or biweight function these two approaches lead to two different situations. Elementwise weighting is more flexible, since small weights could be assigned to separate scores of an object, leaving its other scores unaffected. On the other hand rowwise weighting might conceptually (and computationally) be more attractive, since it considers whole objects as possible outliers. The latter is suitable in the context of PCA, which usually considers a data matrix of objects and variables. The elementwise approach is conceptually more suitable in the bilinear analysis of tables, in which rows and columns play a more symmetric role.

7. Extensions

In this section some extensions are discussed of the general minimization problem formulated in (4). We have seen that (4) can be solved by weighted cyclic dyadic fitting procedures, when there are no restrictions. Now suppose $\mathbf{Z}$ is a matrix where the columns represent categorical variables for which we may assume that they are measured on an ordinal or nominal level. The additional objective is to find optimal transformations of the variables, where optimal is defined in terms of the loss function (Young, 1981). The optimally transformed variables are denoted as $\mathbf{q}_j (j = 1, \ldots, m)$. This yields, as the nonlinear variant of (4), the following function:

$$\xi(\mathbf{Q}, \mathbf{X}, \mathbf{A}) = \sum_{j=1}^{m} (\mathbf{q}_j - \mathbf{X}\mathbf{a}_j')' \mathbf{V}_j (\mathbf{q}_j - \mathbf{X}\mathbf{a}_j'), \tag{26}$$

which has to be minimized over $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{q}_1, \ldots, \mathbf{q}_m$, satisfying $\mathbf{q}_j'\mathbf{q}_j = \mathrm{n}$ and $\mathbf{q}_j \in \Gamma_j$, where Γ_j indicates the set of admissible transformations of the given variable $\mathbf{z}_j$. The class of transformations may be defined differently for each variable, and includes nominal, monotonic, and linear transformations. In fact, (26) is a generalization of the PRINCIPALS program (Young et al., 1978), in which $\mathbf{V}_j = \mathbf{I}$ for all $j = 1, \ldots, m$.

Up to some irrelevant constants, the function in (26) can still be seen as a majorizing function for one of the resistant functions. Minimizing (26) is of course more complex than minimizing (4), but they both lead to a minimum of an objective function (see Verboon et al., 1991). It follows that from a technical point of view optimal scaling can easily be added to the problem of finding a low-dimensional representation of a data matrix.

Another extension is the possibility to deal with missing values in the data. The missing values are not part of the minimization problem by weighting them with zero. For this we need a set of binary diagonal matrices $\mathbf{M}_j$, indicating missing values by 0 and those observed by 1. Each regression problem is now defined in the metric $\mathbf{M}_j$. It follows that the quadratic part of the majorizing function to be minimized becomes:

$$\sum_{j=1}^{m} (\mathbf{z}_j - \mathbf{X}\mathbf{a}_j')' \mathbf{M}_j \mathbf{V}_j (\mathbf{z}_j - \mathbf{X}\mathbf{a}_j'). \tag{27}$$

Obviously, the matrices $\mathbf{M}_j$ are fixed throughout the algorithm. When there are no missing values, $\mathbf{M}_j = \mathbf{I}$ for all variables, and (27) equals (4).

8. Discussion

In the present paper a very general approach based on majorization has been discussed for fitting lower rank approximation of matrices. Many different weight functions can be applied instead of (15) or (16), provided that they yield a proper majorizing function, which guarantees monotonic convergence. Examples are Hampel's three-part hard redescender (Hampel, 1968), Eilers' soft redescender (Eilers, 1987), or simply a trimming function, which assigns 0 to residuals larger than a particular value and 1 otherwise. In all cases, the whole problem is repeatedly applying the algorithm proposed by Gabriel and Zamir (1979), since the choice of the weights yields no problems.

References

Beaton A.E. & Turkey, J.W. (1974), The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, **16**, 147–185

Eilers, P.H.C. (1987), Adaptieve gewichten, een exploratieve techniek voor uitbijters en mengsels van regressie modellen. *Kwantitatieve Methoden*, **23**, 63–83.

Gabriel, K.R. (1971), The biplot-graphic display of matrices with application to principal components analysis. *Biometrika*, **58**, 453–467.

Gabriel, K.R. & Odoroff, L. (1984), Resistant lower rank approximation of matrices. In: E. Diday et al. (Ed.), *Data Analysis and Statistics III* (pp. 23–30). Amsterdam: North-Holland.

Gabriel, K.R. & Zamir, S. (1979), Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, **21**, 4, 489–498.

Good, I.J. (1969), Some applications of the singular value decomposition of a matrix. *Technometrics, 11*, 823–831.

Hampel, F.R. (1968), *Contributions to the theory of robust estimation*. Ph. D. thesis, University of California, Berkeley.

Hawkins, D.M. & Fatti, L.P. (1984), Exploring multivariate data using the minor principal components. *The Statistician*, **33**, 325–338

Heiser, W.J. (1987), Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis*, **5**, 337–356.

Huber, P.J. (1964), Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Huber, P.J. (1981), *Robust Statistics*. New York: Wiley.

Jolliffe, I.T. (1986), *Principals component analysis*. New York: Springer-Verlag.

Verboon, P. & Heiser, W.J. (1992), Resistant orthogonal Procrustes analysis. *Journal of Classification, 9*, 237–256.

Verboon, P., Van der Lans I. & Heiser, W.J. (1991), *The multipals algorithm*. Research Report 91-05. Leiden: Department of Data Theory.

Young, F.W. (1981), Quantitative analysis of qualitative data, *Psychometrika*, **46**, 347–388.

Young, F.W., Takane, Y. & De Leeuw, J. (1978), The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. *Psychometrika*, **43**, 279–281.